DOCUMENT RESUME

ED 474 259 EA 032 339

AUTHOR	Krueger, Alan B.; Zhu, Pei
TITLE	Another Look at the New York City School Voucher Experiment.
PUB DATE	2003-01-00
NOTE	46p.; Financial support provided by the Princeton University Industrial Relations Section. Paper presented at the Conference on Randomized Experimentation in the Social Sciences (Yale University, NJ, August 20, 2002).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Academic Achievement; *Educational Vouchers; Elementary Secondary Education; Racial Factors; *School Choice
IDENTIFIERS	*New York

ABSTRACT

This paper reexamines data from the New York City school-choice program. Data were collected in spring 1997 and the following 3 years from low-income public-school students who were eligible for a private-school scholarship and from their parents. Students with missing baseline test scores, which includes all those who were initially in kindergarten and 11 percent of those initially in grades 1-4, were excluded from previous analyses of achievement. Including students with missing baseline test scores increases the sample size by 44 percent. For African American students, the only group to show a significant, positive effect of vouchers on achievement in past studies, the difference in average followup test scores between the treatment group (those offered a voucher) and control group (those not offered a voucher) becomes statistically insignificant at the .05 level and much smaller if the full sample is used. Also, the effect of vouchers is found to be sensitive to the particular way race/ethnicity was defined. For example, if children with a black, non-Hispanic father are added to the sample of children with a black, non-Hispanic mother, the effect of vouchers is smaller and statistically insignificant at conventional levels. (Contains 17 references and 7 tables.) (RT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 474 259

Another Look at the New York City School Voucher Experiment.

Alan B. Krueger
Pei Zhu

January 2003

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

A. Krueger

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

EA 032 339

BEST COPY AVAILABLE

Another Look at the New York City School Voucher Experiment

Alan B. Krueger
Princeton University and NBER

and

Pei Zhu
Princeton University

January 2003

*We thank Daniel Mayer, David Myers and Christina Tuttle for answering many of our questions and providing data. We have also benefited from helpful comments from Joshua Angrist, David Card, Alan Gerber, Donald Green, Jean Grossman, Bo Honore, Larry Katz, Jens Ludwig, Lisa Markman, Derek Neal, Cecilia Rouse, Jack Stenner and seminar participants at Duke, the University of Virginia and Yale. Financial support was provided by the Princeton University Industrial Relations Section. This paper was prepared for the "Conference on Randomized Experimentation in the Social Sciences," Yale University, August 20, 2002. The usual disclaimer applies.

EA 031 339

Another Look at the New York City School Voucher Experiment

ABSTRACT

This paper reexamines data from the New York City school choice program, the largest and best implemented private school scholarship experiment yet conducted. In the experiment, low-income public school students in grades K-4 were eligible to participate in a series of lotteries for a private school scholarship in May 1997. Data were collected from students and their parents at baseline, and in the Spring of each of the next three years. Students with missing baseline test scores, which encompasses all those who were initially in Kindergarten and 11 percent of those initially in grades 1-4, were excluded from previous analyses of achievement, even though these students were tested in the follow-up years. In principle, random assignment would be expected to lead treatment status to be uncorrelated with all baseline characteristics. Including students with missing baseline test scores increases the sample size by 44 percent. For African American students, the only group to show a significant, positive effect of vouchers on achievement in past studies, the difference in average follow-up test scores between the treatment group (those offered a voucher) and control group (those not offered a voucher) becomes statistically insignificant at the .05 level and much smaller if the full sample is used. In addition, the effect of vouchers is found to be sensitive to the particular way race/ethnicity was defined. Previously, race was assigned according to the racial/ethnic category of the child's mother. If children with a Black (non-Hispanic) father are added to the sample of children with a Black (non-Hispanic) mother, the effect of vouchers is smaller and statistically insignificant at conventional levels.

Alan B. Krueger
Economics Department
Princeton University
Princeton, NJ 08544
(609) 258-4046
akrueger@princeton.edu

Pei Zhu
Economics Department
Princeton University
Princeton, NJ 08544
(609) 258-5694
pzhu@princeton.edu

Now that the Supreme Court has ruled in the *Zelman* case that public funds may be used to support vouchers to enroll children in private religious schools, many states, school districts and parents will seriously consider the desirability of school vouchers. This decision naturally depends on many factors, not least of which is whether school vouchers are likely to raise student achievement. The best currently available evidence on the effect of school vouchers on students' performance is from a series of three randomized experiments conducted in Washington, D.C., Dayton, OH and New York City by Paul Peterson and his collaborators. This paper reexamines evidence from the New York City voucher experiment, which was conducted by Mathematica Policy Research and the Program on Education Policy and Governance at Harvard University.

The New York City experiment was selected because it is the only one of the three experiments for which data have been made available to outside researchers. Two additional reasons argue for a detailed evaluation of the New York experiment, however. First, the New York experiment is the best documented of the three experiments, and had the lowest attrition rate, highest voucher take-up rate, and largest sample size. Second, New York is the only one of the three cities to show significant gains in test scores for voucher recipients relative to non-recipients for African American students at the conclusion of the experiment.[1] In all three experiments, there is no significant difference in student performance between those offered a voucher and the control group for other racial and ethnic groups, or overall.

[1] See Howell and Peterson (2002), Table 6-3. In Washington, vouchers had a statistically insignificant, negative effect on Black students' scores after three years; in Dayton, vouchers had a positive effect that was statistically insignificant at the 0.10 level holding constant family background controls (but significant at the 0.10 level without family background controls) in the second and final year of the experiment.

BEST COPY AVAILABLE

The New York City school choice experiment worked as follows.[2] In February 1997, the School Choice Scholarships Foundation (SCSF), a private foundation, offered 1,300 scholarships worth up to $1,400 a year for three years to children from low-income families (i.e., qualified for free lunch) who were enrolled in Kindergarten through fourth grade in New York City public schools. Some 11,105 eligible students applied for a scholarship between February and late April 1997. Recipients were selected in a series of lotteries in May 1997, and began attending private schools the next fall. Mathematica randomly selected the students offered vouchers (subject to the SCSF requirement that 85 percent of recipients be from public schools in the bottom half of the city-wide test score distribution) and a control group from the eligible applicants. About three quarters of the students who were offered vouchers used them in at least one year; these students overwhelmingly attended religious schools.[3] Information from the students and their parents was collected prior to the lottery and in the spring of each of the ensuing three years. Base weights were constructed so the students in the sample were representative of the pool of eligible applicants (which had 70 percent of students from schools with below-median scores), and the weights were subsequently adjusted for attrition each year.

Students were given the Iowa Test of Basic Skills (ITBS) at baseline and in the spring of each of the three follow-up years. A decision was made not to test the cohort of Kindergarten students applying for scholarships for first grade at baseline, however. (Henceforth, the five cohorts of students in different grades will be referred to by their grade level at baseline.) The Kindergarten cohort was nonetheless given follow-up ITBS tests, along with other students, when they were in grades 1, 2 and 3. In addition, about

---

[2] See Mayer, Peterson and Myers, et al. (2002) and Hill, Rubin and Thomas (2000) for further details.
[3] Very few of the controls attended private school.

11 percent of students initially in grades 1-4 lacked baseline scores.[4] These students were also given the ITBS in the three follow-up waves. The sample weights do not attempt to adjust for missing baseline scores.

A contribution of our paper is that we include students with missing baseline scores in much of our analysis. Previous analyses of achievement omitted students with missing baseline test data.[5] Howell and Peterson (2002), for example, report, "A handful of additional families were offered vouchers, but they were not included in the evaluation for lack of baseline [test] information." Because of random assignment, however, estimates are unbiased even without conditioning on baseline information, so there is an efficiency loss from excluding these students entirely. For the subsample with baseline scores, omitting the baseline score only trivially affects the estimated treatment effect, as one would expect with random assignment. Including students with missing baseline test data increases the sample size by 44 percent in the third and final follow-up year; nearly 30% of those with missing baseline scores were in grades 1-4 when the experiment started and should have had baseline scores. An argument can be made that including students with missing baseline scores – both those in the Kindergarten cohort and those in the other cohorts – is desirable because the weights make no provision for sample exclusion due to missing baseline scores, and, more importantly, because using a sample that encompasses more grade levels enhances the generalizability of the results.

---

[4] This is in addition to students who had scores of 0 on the baseline test – of the 1,851 students with baseline scores, 199 (10.7%) received a zero score in reading and 324 (17.5%) received a zero score in math; 97 (5.2%) received a zero on both exams.

[5] When Mayer, Peterson and Myers, et al. (2002) examine outcomes such as parental satisfaction, however, they include observations on students without baseline test data. It is unclear whether Howell and Peterson (2002) include or exclude students with missing baseline tests when they study parental responses. Their Table 2-3 reports that they include students entering grades 1-4, but that apparently is a misprint, as their sample includes students entering grades 2-5. Nevertheless, we could only replicate some of their results based on the parental survey if we include the Kindergarten cohort.

BEST COPY AVAILABLE

For African American students, the estimated effect of being offered a voucher is much weaker if students with missing baseline scores are included in the analysis. In the third follow-up year, for example, the estimated effect of being offered a voucher on composite test scores is 2.78 percentile points (t = 1.65) if baseline test scores are dropped from the model and the larger sample is used; controlling for baseline covariates other than test scores, the effect is 2.12 points with a t-ratio of 1.27.[6] For comparison, Mayer, Peterson and Myers report a coefficient of 5.5 points with a t-ratio of 3.42. Results are also weaker in the first two follow-up years if students with missing baseline scores are included in the sample. These findings raise doubts about the robustness of earlier findings of a significant positive effect of offering vouchers on test scores for African American students.

In the next section, we discuss and evaluate the random assignment procedures in more detail. In the following section we explore the sensitivity of the results to including and excluding students without baseline scores, and examine differences in the treatment effect across cohorts.

Although results for the larger sample encompassing students with and without baseline scores cast doubt on the inference that vouchers had a positive impact on Black students' test scores, we explore reasons why vouchers might have been more effective at raising scores for Black students than for other students in the grade 1-4 cohorts. Results presented in Section 4 suggest that differential characteristics of the initial public schools Black and non-Black students attended are *not* responsible for any differential effect by race, even in the sample analyzed by Howell and Peterson.

---

[6] These estimates control for 30 dummies indicating the original strata students were placed in for random assignment. The definition of the strata is described in detail below.

Indeed, the data suggest that race itself independently affects the gain from vouchers. This leads us to closely examine the particular definition of race used in the experiment. There is no universally accepted definition of race. Mathematica assigned students to a racial/ethnic group based on a single question on the parental survey that asked respondents to *select only one* of the following categories for the mother or female guardian: Black/African American (non-Hispanic), White (non-Hispanic), Puerto Rican, Dominican, Other Hispanic, etc. Students were assigned the mother's race/ethnicity, regardless of the father's race or ethnicity. Thus, if a student is reported as having a Black mother and a Hispanic father, he or she was classified as Black. If the same student had a Black father and Hispanic mother, however, he or she would have been classified as Hispanic. Arguably, father's race is also relevant. If we augment the sample to include students for whom *either* parent is classified as Black/African American (non-Hispanic), the effect of offering a voucher on the composite score in year three falls to 1.52 points with a t-ratio of 1.04. So in the broader sample the estimated effect of offering vouchers to students with a Black parent is small and statistically insignificant.

Throughout the paper, we focus mainly on intent-to-treat estimates; that is, the impact of offering students a voucher on their test performance, as opposed to the effect of *attending private school* on test performance. We focus on intent-to-treat estimates because offering a voucher – as opposed to compelling students to switch to private school -- is the policy decision that is most relevant, and because there is a cleaner statistical interpretation of the intent-to-treat estimates in this case. Nevertheless, the effect of attending a private school for varying lengths of time is also of interest. In Section 5 we present Instrumental Variables results that estimate the impact of the

number of years in private school on student achievement. These results differ from those emphasized by Howell and Peterson (2002), who examine the effect of attending private school for three full years, and implicitly make strong assumptions about the effect of switching to private school on achievement for those who attend private school for fewer than three years.

1. Randomization

The procedures used to randomly assign students to treatment and control status, and select control group members for follow up, are described in Hill, Rubin and Thomas (2000). Because multiple children from many families applied for scholarships, and it was desired to assign all family members to the same treatment status, students were assigned to control and treatment groups in a lottery in which *families* were the unit of observation. Two methods of random assignment were used.

Briefly, for students from 1,000 families a Propensity Match Pairs Design (PMPD) was used, and for students from 960 families a Stratified Block design was used. The PMPD method, which introduces considerable complexity to the design, was used because in the first lottery many more potential control group members were available to be followed up than was money to follow them up. Rather than select a random sample of the controls to follow up, it was decided that it would be more efficient to follow up the subset that, in some sense, is most alike to treatment group members. Consequently, after the treatment group was randomly selected, the members of the control group who were followed up were selected by estimating a propensity score model to identify those with attributes that were closest to members of the treatment group.

According to Hill, Rubin and Thomas (2000), variables used in this model included, in order of importance: family size, a dummy indicating above versus below median schools, grade level, and initial test scores.[7] Students with missing data were also included in the selection. Once the propensity score model was estimated, a matched pair for each treatment group family was selected by choosing the nearest available neighbor Mahalanobis match from among those with propensity scores close to that of each treatment group member.

The Stratified Block design was much simpler. Samples of screened applicants were invited to participate in four sessions at which baseline data were collected and ITBS tests were administered. In these lotteries, by design approximately 85 percent of the invitees were from schools with below the city median test score. Treatments and controls were randomly assigned in four lotteries, and a random sample of participants were included in the follow-up sample. Each of these lotteries constitutes a block.

One can define 30 mutually exclusive "random assignment strata": 5 lottery blocks (PMPD block plus 4 stratified blocks) x 2 school types (above and below city median test) x 3 family size groups (1, 2 or 3 or more students). Assignment is random within these strata. After the lotteries were held, Mathematica discovered that some families had misreported their family size and were thus placed in the "wrong" strata; revised strata were created with the latest family size information. Nevertheless, assignment to treatment status is random within the *original* strata that were actually used to apportion the sample at the time of random assignment. Peterson and Howell (2002)

---

[7] Other variables included, in order of importance, were: ethnicity, mother's education, participation in special education, participation in a gifted and talented program, language spoken at home, welfare receipt, food stamp receipt, mother's employment status, educational expectations, number of siblings, and an indicator for whether the mother was foreign born.

and Mayer, Peterson and Myers, et al. (2002), however, controlled for dummies indicating membership in the revised strata, not the actual strata used to make assignments. Unless otherwise specified, our results condition on the *actual* strata that were used when random assignments were made. Fortunately, as shown below, the choice of original or revised strata has relatively little impact on the results.

Because assignments were over families, not students, and children from the same family tend to have correlated outcomes, we compute bootstrap standard errors that use families instead of individual students as the unit for resampling to allow for dependence across family members.[8] (Forty-six percent of students in the sample had at least one sibling in the sample as well.) Moulton (1990) provides a nice illustration of inference problems that can arise from ignoring correlated errors.

Table 1 reports the mean of several baseline characteristics for the treatment and control groups for the full sample, separately for Black and Latino students, and disaggregated by cohort for Black students. Because random assignment was implemented within strata, regressions were estimated to condition on the 30 original randomization strata, and conditional treatment-control differences and t-tests are reported as well. For the overall sample, the results indicate small and statistically insignificant differences between the treatment and control groups. For example, the conditional t-test for the difference in the mean baseline composite test score between treatments and controls is only -0.63. Mayer, Peterson and Myers, et al. (2002) present similar results. Hill, Rubin and Thomas (2000) report treatment-control differences by PMPD versus Stratified Block strata, and find slightly better balance in the PMPD strata, but in both cases there are not systematic differences between the treatments and controls.

In the first-year follow-up report, however, Peterson, Myers and Howell (1998) reported highly statistically significant differences in *baseline* test scores between the treatment and control group, with control group members scoring significantly higher on both the math and reading exams. Evidently, this was a result of inaccurate weights. The baseline weights did not adjust correctly for the size of the underlying assignment strata.[9] In the subsequent reports, the baseline weights were revised "to include post-stratification adjustments," and the baseline differences in test scores were no longer statistically significant. The complexity of the design of the experiment probably contributed to the initially inaccurate baseline weights, as the average discrepancy between the initial and corrected baseline weights was 316 percent for the PMPD strata and only 24 percent for the Stratified Block strata. [10]

The results in Table 1 suggest that the assignment groups were well balanced, as one would expect with random assignment. One exception, however, is the oldest cohort of African American students. For this group, at baseline the treatments' mean score is higher than the controls', and the treatments are more likely to come from higher income families with better educated mothers. For African Americans as a whole, however, the differences in baseline characteristics are typically insignificant.

1.1 Precision

The PMPD design was intended to reduce sampling variance and improve the balance between treatments and controls. Did the precision of the estimates improve?

---

[8] The sample was drawn with replacement 10,000 times to compute the bootstrap standard errors.
[9] Email communication from David Myers of Mathematica, February 5, 2001.
[10] Just before the completion of this draft, we were informed by David Myers that, as a result of one of our queries, Mathematica had discovered a problem with the revised baseline weights that affected a small number of cases. Mathematica plans to issue new weights in the future. We plan to utilize the revised, revised weights when they become available.

BEST COPY AVAILABLE

Table 2 reports separate estimates of the treatment effect and associated standard errors for the two methods of random assignment. The treatment effect in this instance is the average impact of being offered a voucher on students' test scores, measured in National Percentile Ranks (NPR's). Formally, the treatment effect, • ; is computed by estimating the following regression model for each follow-up year:

(1) $$Y_{if} = \bullet\bullet Z_f + \bullet \downarrow_{if} + \bullet \pi,$$

where Y_{if} is the composite test score (that is, average NPR on reading and math), Z_f is a dummy variable that equals one if the student was offered a voucher and zero if not, $\bullet \downarrow_{if}$ is a fixed effect for the randomization strata (j = 1, ..., 30), and $\bullet \pi$ is an error term that is possibly correlated among members of the same family. The subscript i indexes students and f indexes families. The strata fixed effects are controlled for by including 30 dummy variables indicating each original randomization stratum.

We also present another set of estimates in which the regression model is augmented to control for baseline math and reading test scores, using as a sample the subset of students who have baseline test scores. Formally:

(2) $$Y_{if} = \bullet\bullet Z_f + \bullet\bullet_M M_{if} + \bullet\bullet_R R_{if} + \bullet \downarrow_{if} + \bullet \pi,$$

where M is the math NPR score, R is the reading NPR score, and the • •s are coefficients.

Estimates of • •for the full sample are in Panel A, and for African Americans in Panel B. The rows of the table indicate whether the underlying regression model controlled for baseline test scores or omitted baseline scores and used a larger sample. The second to last column reports the standard error in the Stratified Block design relative to the standard error in the PMPD design. Because the sample sizes differ

slightly, in the last column the relative standard errors are scaled by the ratio of the square root of the relative sample sizes.[11]

For the full sample, in most cases the standard error of the estimated treatment effect is about 10 percent larger in the Stratified Block design than in the PMPD design. The adjustment for sample size has little effect on this ratio. The treatment effect is insignificantly different from zero under either design. Thus, there is a small gain in power from the PMPD design.

Unexpectedly, for the subsample of black students the standard errors are not consistently smaller in the PMPD design than in the Stratified Block design. This result is surprising because, as Hill, Rubin and Thomas point out, analyses of subgroups are expected to have more power in the PMPD design because matching should lead to a more equal representation of subgroups in the treatment and control groups.

One could question whether the gain in precision from the PMPD random assignment procedure was worth the cost of added complexity. For practical purposes, the difference in precision between the two designs is fairly small, even in the full sample: in year 3, for example, the width of a 95% confidence interval increases trivially, from +/-3.1 points in the PMPD design to +/-3.5 points in the Stratified Block design. If the complexity of the experimental design contributed to the incorrect initial inference about the baseline difference in test scores (because of inaccurate weights), then in this case the PMPD design would seem to have been hardly worth the gain in power. This likelihood seems to us to underscore Cochran and Cox's (1957) advice: "A good working rule is to use the simplest experimental design that meets the needs of the occasion."

---

[11] At baseline, the sample size was 1,341 for the PMPD sample and 1,325 for the Stratified Block sample. The sample sizes in Table 2 differ because of attrition and missing follow-up test data. If attrition is

2. Sensitivity of Earlier Results

Next we explore the sensitivity of the estimated impact of offering a voucher to controlling or not controlling for baseline scores, and to including the cohort of students who were in Kindergarten when the experiment began. The first set of columns in Table 3a report estimates of equation (2): a regression of the test score NPR on a dummy indicating whether the student was randomly selected to be offered a voucher, baseline test scores, and 30 dummy variables indicating the actual stratum the family was allocated to for random assignment. The coefficient ($\bullet$) on the voucher offer dummy, also known as the intent-to-treat (ITT) estimate, is presented in the table, along with its standard error and t-ratio. Because it is useful to compare our results to previous studies, in Table 3b we report corresponding results for models that control for the revised randomization strata. To exploit the experiment, it is more appropriate to control for the strata that were used at the time of random assignment, although the results are not very different in either case.

The first set of results in Table 3b exactly replicate those in Mayer, Peterson and Myers, et al., Tables 20-22. As they found, for Black students the results indicate significantly higher test scores for those offered a voucher than for the controls. For Latino students the voucher effect is negative but statistically insignificant. The results are very similar in Table 3a, which condition on the original randomization strata.

The middle set of columns report estimates for models that use the same sample of students – i.e., those who were in grades 1-4 *and* have baseline test data – but omit the baseline test scores (i.e., equation (1)). Omitting the baseline scores has a trivial effect on

---

endogenous to the design, then the unadjusted comparison is more appropriate.

the ITT estimate, as one would expect given random assignment. For the year three composite score for African Americans in Table 3a, for example, the estimated treatment effect is 5.2 points controlling for baseline scores and 5.1 points without controlling for baseline scores. Similar stability is found in Table 3b, with the revised strata controls.

The third set of columns report estimates of the same model for the largest possible sample, including students initially in Kindergarten (all of whom were not tested at baseline) and those initially in grades 1-4 with missing baseline scores.[12] Seventy-one percent of the additional observations are from the Kindergarten cohort, and 29 percent from the other cohorts. The results for African American students are notably weaker in the larger sample. In year three, for example, the estimated treatment effect in Table 3a falls about in half, to 2.78 points (s.e. = 1.68), with a t-ratio that is about equal to the threshold value for statistical significance at the 0.10 level.[13] By subject, the effect on the reading test is even smaller, while the treatment effect for the mathematics test is larger and statistically significant at the 0.05 level.

The notably smaller estimates when students with missing baseline scores are included are primarily due to the inclusion of students who were in Kindergarten at baseline, not those who were in grades 1-4. If the Kindergarten cohort is excluded from the sample in the model in Table 3a that omits baseline scores, the treatment effect in the composite score regression for African Americans is 4.56 (s.e. = 1.83; N = 577).

Interestingly, the standard errors are only trivially different between the estimates that condition on baseline test scores and those that do not but use a larger sample. In the

[12] Note that the cohort in Kindergarten at baseline was in grade 1, 2 or 3 during the follow ups, so testing them in the follow-up waves presumably did not pose greater challenges than testing the other cohorts when they were in those grades.

third follow up, for example, the standard error for the treatment effect on African Americans' reading scores is 0.05 points smaller using the full sample and omitting baseline scores than it is in the subsample that controls for baseline scores, and for math scores it is 0.02 points larger when baseline scores are omitted. The reason for this result is that the reduction in residual variance from controlling for baseline scores is roughly offset by the gain in precision from having a larger sample.

One can improve the precision of the estimates in Table 3 by controlling for baseline information other than test scores. Table 4 presents results from models that control for the original 30 randomization strata, cohort (i.e., grade at baseline), gender, log family income, mother's years of schooling, indicators for whether the mother works full- or part-time, an indicator for whether the student has had special education, an indicator for whether the student has been in a gifted or talented class, student's age, and dummies for English spoken at home, mother born in the United States, family welfare receipt, mother resided in residence for more than one year, and whether the mother's religion was Catholic.[14] The first set of columns present results from models that also control for baseline test scores, for the subsample of observations with baseline scores.

Results shown in the second set of columns are based on the larger sample and omit baseline test scores from the model. The treatment effect of vouchers is smaller in these models. In year 3, for example, the impact of offering a voucher on the composite score for African Americans is 2.12 NPR's (t-ratio = 1.27). Again, the impact on math scores is larger than the impact on reading scores.

---

[13] The conventional standard error in this case is 1.66, not very different than the one we compute, which allows for within family correlated errors.

Finally, it is possible to incorporate the information on baseline scores for those who have it, and still use the larger sample. In particular, we created a dummy variable that equals one if baseline scores are available and zero if they are missing.[15] Define this variable as D. We then interacted D with the baseline math and reading scores, which imposes a value of zero for those with missing scores regardless of the values one would have used to impute those scores. The following "hybrid" model was estimated using the full sample of observations, both those with and without baseline scores:

(3) $Y_{if} = \cdot\cdot Z_f + \cdot\cdot_M M_{if} \bullet D_{if} + \cdot\cdot_R R_{if} \bullet D_{if} + \cdot\cdot D_{if} + \mathbf{X}_{if}' \cdot\cdot + \cdot\cdot_{if} + \cdot\cdot_{if} + \cdot\cdot_{if}$,

where **X** is a vector of other baseline covariates, such as gender, mother's education and log family income, and $\cdot\cdot$ is a set of cohort fixed effects. This model has the effect of controlling for baseline scores for students who have them, and not controlling for baseline scores for student who lack them. Because baseline scores can be missing for two distinct reasons – either students are in the Kindergarten cohort and not tested by design or they are in the other cohorts and failed to comply with testing – controlling for cohort dummies and D in the hybrid model removes differences between these groups.

Results of estimating $\cdot\cdot$ from equation (3) without **X**-covariates (but with cohort and lottery strata dummies) are reported in the third set of columns in Table 4, and with additional **X**-covariates in the fourth set of columns of the table. Notice that the standard errors from both hybrid models are smaller than those from the other models in the table. The hybrid model with the additional **X**-covariates in the right-most set of columns, it turns out, typically yields the most powerful estimates of the models we have considered.

---

[14] So as to not reduce the sample size, missing values of income and mother's education were replaced by their means, and dummy variables were included indicating whether income, mother's education, students' gender, mother's employment status, and each of the other covariates were missing.

[15] It is never the case that math scores are missing and reading scores are available, or vice versa.

BEST COPY AVAILABLE

Controlling for baseline test scores when they are available results in a slightly higher estimated effect of vouchers when the **X**-covariates are excluded. In the model without the **X**-covariates, the year three impact of offering a voucher on the composite test scores of African American students is just below the threshold for statistical significance at the 0.05 level. In the model with additional baseline covariates – which yields a slightly more precise estimate -- the effect is somewhat smaller, 2.19 NPR's, and not statistically significant at the 0.10 level. This point estimate is less than half as large as the estimate in Mayer, Peterson and Myers, et al. (2002), and more precisely estimated.

2.1 To Control, or Not Control

Mayer, Peterson and Myers, et al. (2002, p. 9) give the following justification for why they controlled for baseline test scores: "Baseline characteristics were included to adjust for chance differences between the characteristics of treatment- and control-group members and to increase the precision of the estimated impacts." But because both sets of estimates -- those with or those without baseline characteristics – give an unbiased estimate of the treatment effect, only the latter rationale is potentially valid.[16] Any improvement in the estimates from controlling for covariates is *fully* reflected in the sampling variance of the estimates.

The key question then is whether including baseline covariates reduces the residual variance by enough to justify the loss of degrees of freedom and the fact that, in practice, some observations must be dropped from the sample because they have missing baseline data. On this basis, there is little reason to chose between the estimates in Table

[16] Another way to see this is to note that if there is a chance difference in a baseline characteristic between treatments and controls, there could also be an erroneous correlation (due to chance or misspecification) between the baseline characteristic and the outcome variable that would sway the estimates if covariates are

BEST COPY AVAILABLE

3a that control for baseline test scores and the ones that do not. Both are unbiased with about equal power. However, an advantage of estimates using the larger sample is that the results potentially can be generalized to a broader population – namely, students initially in grades K-4, as opposed to those in grades 1-4 without missing baseline data. Moreover, estimates without baseline covariates are simple and transparent.

Lastly, notice that equation (3) in the far right of Table 4, which incorporates the baseline covariates for those who have them and still uses observations for those with missing baseline data, provides the most precise estimates and pertains to grades K-4. An argument can be made that, of the estimates presented so far, the most credence should be placed in these estimates. As we explain in Section 4 below, however, we think the sample of Black students used in Table 4 is still unnecessarily restrictive.

2.2 Cohort Interactions

The results in Tables 3a,b and 4 suggest that including students who were enrolled in Kindergarten at baseline in the sample qualitatively weakens the impact of school vouchers on achievement scores of African American students. These students were in grades 1-3 when the follow-up tests were administered (assuming they were not held back), and were among the youngest students in the experiment. One possibility is that vouchers are effective for older African American students, but not younger ones. Mathematica (2000) raised this concern after the second-year follow-up study. Howell and Peterson (2002; p. 222) strongly argue against this interpretation, however, maintaining that "no decipherable pattern suggests that impacts vary by grade level." Our analysis of inter-cohort differences in treatment effects also suggests that the grade at

---

included. A correlation between treatment status and baseline covariates in an experiment is potentially problematic regardless of whether baseline covariates are controlled in a regression.

which students are offered vouchers is unrelated to the magnitude of the treatment effect in the third year of the experiment when we test the null hypothesis of any unrestricted pattern of cohort effects, although we find some tendency for older students to have a larger treatment effect when Kindergarten students are included in the sample and the cohort-treatment-status interaction is constrained to be linear.

The treatment effect on the third-year composite test score for each cohort can be viewed in the Addendum at the bottom of Table 1. The treatment effects are not uniform, and it is particularly large for the 4th grade cohort, but they individually have large standard errors. To conduct a formal test for inter-cohort variability in treatment effects, we interacted five dummy variables indicating membership in one of the initial grade cohorts with the treatment status dummy in the models in Table 4, and conducted an F-test of the null hypothesis that the treatment effects were equal for all cohorts. In the third follow-up year, there was no more than a chance occurrence that the treatment effect differed across the cohorts for African American students. For example, in the model for the composite score in the far right of Table 4, the p-value for an F-test of uniform treatment effects is 0.21. This F-test puts no structure on the pattern of interactions and does not depend on first screening the pattern of coefficients to test a particular alternative hypothesis, so in this sense it is "hands above the table."

Observing the pattern of treatment effects by cohort, however, it appears that there is a tendency for the older cohorts to have larger coefficients than younger cohorts. If we perform a simple test for differential cohort effects that recognizes the *order* of the cohorts by including an interaction between a student's linear grade level at baseline (0, 1, 2, 3, 4) and the voucher offer dummy, as well as the voucher offer main effect, the p-

value for a test of the interaction term in the most powerful model in Table 4 is 0.05 for the composite score. Thus, there is some evidence that the treatment effect rises with the entry grade of the cohort of African American students in the full sample, but it is not overwhelming and it requires the imposition of a linear cohort-treatment interaction. Moreover, it is unclear why the treatment effect would vary with age for African American students but not Hispanic students.

2.3 Attrition

One commonly voiced criticism of the voucher experiments is that attrition was high, especially for the control group (e.g., Neal, 2002). The response rate in the third year of the New York experiment was 68.8 percent for treatment group members and 64.6 percent for control group members in cohorts 1-4. The rates were slightly higher for the Kindergarten cohort: 71.7 percent for treatments and 69.2 percent for controls. Although, in principle, the follow-up sample weights adjust for attrition based on observable characteristics, it is possible that those who exited the sample differ along unobserved dimensions, and the average achievement in year three for non-responding treatments and controls could be different than it was for respondents. Nonrandom attrition that is correlated with treatment status would bias the weighted estimates.

In results not reported here, we examined the impact of attrition in a couple of different ways. First, we computed two-step normal selection correction estimates to adjust for those with missing follow-up test scores. Second, we used the technique employed in Krueger (1999) of carrying forward the last available test score for those with missing data. We also used these techniques to adjust for students with scores of zero. These adjustments typically led to only slightly smaller estimates of the effect of

vouchers. Thus we conclude that nonrandom attrition was probably not a major problem in the experiment.

3. The Relevance of Race and Residence

Our interpretation of the results so far is that the impact of vouchers on achievement for Black students in New York City is smaller and less robust than has been previously acknowledged. Nevertheless, we examine reasons why there may have been gains using the sample that displayed the largest impact, Black students in the grade 1-4 cohorts with baseline test data.

The inference that African American students benefited academically from scholarships to attend private school would be more credible if there were a compelling explanation for why African American students benefited while other, equally poor Hispanic and White students did not benefit. Howell and Peterson (2002) offer the following *ex post* explanation: "African Americans, more than other groups, live in the poorest, least attractive, and most dangerous communities within metropolitan regions. … Precisely because African Americans suffer most under a system of public education based on residency, they stand to benefit the most from the new education opportunities that vouchers afford."

In short, they hypothesize that constrained residential choices for African American families – but not Dominican or Puerto Rican families – account for the poor performance of public schools attended by African American students. As a result, African American students are disproportionately confined to public schools that under perform vis-à-vis private schools. This is a plausible and potentially quite important

hypothesis, with many striking implications.[17] Indeed, if discrimination in the housing market were responsible for the alleged poorer performance of Black students in public school, then enforcing anti-discrimination laws could possibly be more effective at raising achievement than school reform.

Howell and Peterson's hypothesis is testable. In particular, if differential residential location and therefore differential school attendance explains the results, then the minority of non-Black students who attend the *same* public schools as Black students should benefit from being offered private school scholarships as much as Black students. We find no support for this hypothesis.

To test the hypothesis, we identified all the non-Black students in the sample who, at the time of baseline, attended a public school that was also attended by a Black student in the sample that year. In the third follow-up, a total of 470 non-Black students and 333 Black students in cohorts 1-4 came from overlapping baseline schools. We then adjusted the sample weights for this subsample of non-Black students so that weighted counts would correspond to the (weighted) distribution of Black students in the overlapping schools. Consequently, we can produce results for a sample of non-Black students in the experiment who attended the *same* distribution of schools as a subsample of Black students in the experiment. If differential residence and school attendance patterns by race are responsible for the differential effect of school vouchers, we would expect the sample of non-Black students in this subsample to exhibit the same treatment effect of vouchers as the Black students.

---

[17] One fact that is inconsistent with this hypothesis, however, is that, on average, the Hispanic students have lower baseline test scores and lower family incomes than the Black, non-Hispanic students in the sample.

This was not the case. If anything, the results point in the opposite direction. For third year math scores, for example, the treatment effect of vouchers for the sample of Blacks from overlapping schools is 5.47 (s.e. = 2.37) and for non-Black students is –3.27 (s.e. = 3.26).[18] Even when attention is limited to the set of overlapping schools that Black and non-Black students attended, the non-Black students do not appear to gain from being offered a voucher; indeed, they do worse, and the differential treatment effect is statistically significant, with a t-ratio of 2.17. In addition, we find an insignificant effect if we interact the percent of students in the baseline school who are black (derived from Common Core Data reported by schools) with the treatment status dummy for non-Black or for Black students. These findings suggest that differential characteristics of the initial public school that students with different racial backgrounds attended do *not* account for any gain in test scores that Black students may have reaped from attending private school.[19]

This conclusion is a contrast to Krueger and Whitmore's (2002) findings on racial differences of the effect of class size. They find that the Black students benefited more from attending smaller classes than White students, but when they estimate effects for White students who attended the same mix of schools as Black students, the White students benefited equally. They concluded, "These findings suggest that small classes matter for Blacks because of something having to do with the schools they attend, rather than something inherent to individual Black students *per se*." In the case of New York

[18] These results control for baseline test scores and 30 strata dummies. Similar results arise if reading data are analyzed, or if data from all five cohorts are used and baseline scores are dropped from the model. We focus here on the older cohorts to give the hypothesis a stronger chance of being supported by the data.

[19] From examining administrative data on school characteristics and parental reports on the school environments, Mayer, Myers and Tuttle (2002) similarly conclude, "Differences in school characteristics do not appear to explain the differences in test scores between the African American and Latino controls."

BEST COPY AVAILABLE

City school vouchers, if one accepts that there is an impact for Black students, something about race itself would seem to be inherently part of the reason.

4. Hispanic and Non-Hispanic Black Mothers and Fathers

If race matters, it is important to know how race is defined. Race is a social construct that varies from survey to survey. In Howell and Peterson (2002), Black students are defined as children whose mother or female guardian is reported as Black/African American (non-Hispanic) in the parental survey. Specifically, the Mathematica baseline survey asked respondents to "MARK ONLY ONE" of the following racial/ethnic categories for each parent or guardian: "Black/African American (non-Hispanic); White (non-Hispanic); Puerto Rican; Dominican; Other Hispanic"; American Indian, Chinese, etc. Mathematica then coded a child's race as Black if the mother or female guardian responded "Black/African American (non-Hispanic)". Consequently, a child's race and Hispanic ethnic origin are by definition mutually exclusively in these data. This deviates from the Office of Management and Budget's guidelines (Statistical Policy Directive No. 15), which recommend separate questions for race and ethnicity self-identification questions.[20]

We find from the 1990 Census that in the New York metropolitan area 15 percent of low-income ($\leq$ $20,000) children age 5-14 whose race is reported as "Black, African American or Negro" are also reported as Hispanic.[21] And 28 percent of those who indicate their ethnicity as Dominican indicate their race as Black/African American. A

[20] In other surveys, such as the National Job Corps Study, Mathematica has used the OMB format of asking separate race and ethnicity questions.

[21] This is based on a sample of children living in the New York-Northeastern New Jersey Consolidated Metropolitan Area.

total of 541 students whose race/ethnicity was classified as Dominican participated in the voucher experiment, almost half as many as the number of non-Hispanic Blacks who participated. Many of those who were categorized as Dominican probably would have identified themselves as Black if the Census definition of race had been used in the study. Because the treatment effect of vouchers was negative for students whose race was classified as Dominican Republic, including Black Dominicans in the sample of Black students would likely lead to smaller estimated treatment effects.

Another problem is that the practice of assigning a child the race/ethnicity of his or her mother irrespective of his or her father's race is asymmetric and restrictive.[22] Some countries, such as China, legally assign a child's race based on the father's race, while others use the mother's race, and still others allow families to choose. In part, race may matter because society treats individuals with different skin tones differently (see, e.g., Darity, Hamilton and Dietrich, 2001, Keith and Herring, 1991 and Scarr, et al., 1977). If this is the case, then one could argue that treating mothers and fathers symmetrically is more sensible than assigning race according to the mother's race, regardless of father's race. Moreover, race was reported as Black for 85 percent of the children with a Black father and a Hispanic mother present in the New York metropolitan region according to a tabulation of the 1990 Census. Because most of the children with a Black father who were not classified as Black in the Mathematica sample had a Hispanic (possibly Black) mother, it is quite likely that these students would have been classified

[22] As an example of the types of problems created by this procedure, note that eight students in the data set had a Black (non-Hispanic) father and were missing information on their mother's race; these students were classified as non-Black. In three of these cases, there is no indication the mother lived at home.

as Black had they or their parents been given the opportunity to report the race of the child.

In Table 5 we present results for the group of students for whom either parent's race/ethnicity is identified as Black/African American (non-Hispanic), which increases the sample by about 10 percent. This is a broader definition of Black students' race than the one employed previously, although it still treats race and Hispanic origin as mutually exclusive, contrary to the OMB guidelines. Notably, the results for this sample are weaker than those in the original sample. In the full sample, including students with missing baseline scores, the effect of offering a voucher on the third-year composite score is quite small (1.52 NPR's) and statistically insignificant at conventional levels (t-ratio=1.04). To put this in context, note that the standard deviation of percentile ranks in the national population is 28.9 (because percentiles are uniform), so the estimated effect size is only 0.053 standard deviations. Moreover, if we interact treatment status with three dummies indicating parents' race (mother Black, father Black, both parents Black), we do not reject that the treatment effect is the same for all three groups.[23]

These results cast further doubt on the extent to which one can generalize previous findings from the voucher experiment to the population of low-income Black students. We would stress that the qualitatively different results in Table 5 stem from two plausible changes from previous estimates: students with missing baseline scores are included and parental race is treated symmetrically for assigning children's race. *We regard the results in Table 5 as the most relevant estimates of the impact of offering*

[23] The coefficients (with standard errors in parentheses) for the voucher-race interactions in the third model of Table 5 are: 1.24 (1.72) for both parents Black (non-Hispanic); -1.14 (5.71) for father Black, mother non-Black or Hispanic; and 2.22 (2.43) for mother Black, father non-Black or Hispanic. The p-value for an F-test of the null hypothesis that the three effects are equal is 0.92.

BEST COPY AVAILABLE

private school vouchers on achievement for the broadest sample of Black elementary school children. The finding that the results are so sensitive to these defensible changes in the sample leads us to conclude that the provision of vouchers in New York City probably had no more than a trivial effect on the average test performance of participating Black students. Furthermore, if parents were not constrained to give a mutually exclusive response to their race *and* ethnicity in a single question, contrary to the OMB guidelines, the effect of vouchers for Black students of *any* ethnicity would likely be even smaller.

5. Effect of Time in Private School – Instrumental Variables Estimates

Although the ITT estimates of the effect of offering a student a voucher are of much interest, another question concerns the effect of *attending* private school on student achievement. Howell and Peterson (2002) and Mayer, Peterson and Myers, et al. (2002) report Instrumental Variables (IV) estimates to estimate the impact of attending private school on student achievement. Specifically, they create a dummy variable that equals one if the student attended private school for three years, and zero otherwise. They then use the voucher offer dummy as an instrument for private school attendance. Because the voucher was randomly offered, this approach yields a consistent estimate of the population parameter. (If there are heterogeneous treatment effects across students, additional assumptions are necessary to interpret the parameter as the causal effect on compliers; see Angrist, Imbens and Rubin, 1996.) As Rouse (1997) and Mayer, Peterson and Myers, et al. point out, if switching to private school for one or two years raises (lowers) achievement, this approach will overstate (understate) the impact of attending private school for three years versus not at all. Mayer, Peterson and Myers, et al. also

present estimates where the endogenous regressor is a dummy indicating whether the student ever attended private school.

Intuitively, all the IV estimates involve scaling the ITT estimate emphasized previously by the difference in some quantity between those offered and not offered a voucher. To see this, note that if there are no covariates the IV estimate of the effect of attending private school for three years, β_1 has probability limit:

(4) $\quad \text{plim } \beta_1 = \{E[Y \mid Z = 1] - E[Y \mid Z = 0]\} / \{E[P3 \mid Z = 1] - E[P3 \mid Z = 0]\},$

where Y is test scores, Z is a dummy that equals one if a voucher was offered and zero if not, and P3 is a dummy variable that equals one if the student was enrolled in private school for all three years of the experiment and zero otherwise. The numerator is the ITT estimate. In a model in which attending private school for at least one year is the endogenous regressor, the numerator is the same but the denominator is the treatment-control difference in the probability of attending private school at least one year.

An alternative approach is to treat the number of years in private school as the relevant variable of interest. Assuming years in private school have a linear effect on achievement, the equations of interest are:

(5) $Q_{if} = \beta_0 + \beta_1 Z_f + \mathbf{X}_{if}'\beta_2 + \lambda_{1f} + \pi_i$

(6) $Y_{if} = \beta_0 + \beta_1 Q_{if} + \mathbf{X}_{if}'\beta_2 + \lambda_{1f} + \eta_i,$

where Q_{if} is the number of years the student has spent in a private school up to that point of the experiment, Z_f is a dummy indicating whether the student was offered a voucher, $\mathbf{X}_{if}$ is a vector of baseline covariates, Y_{if} is the test score in the relevant follow-up year, λ_{1f} and λ_{1f} are randomization strata fixed effects, and π_i and η_i are equation errors. Interest is in the parameter β_1, which measures the marginal impact on performance of

attending private school an additional year. An advantage of the linear model is that the coefficient • ϒ can be compared in different years. Furthermore, for comparison to • ;an estimate of the effect of attending private school for all three years is 3• ϒ.

Two-Stage Least Squares estimates of • ϒ using various samples are presented in Table 6. Similar to our earlier findings, the results indicate that spending more time in private school has a positive and statistically significant effect on achievement for children of Black (non-Hispanic) mothers when the analysis is confined to those with baseline test scores, but an insignificant effect in the larger samples. In the third follow-up, the implied effect of attending a private school for three years is 6.7 points (t = 2.75) in the sample with baseline test scores, and 3.5 points (t = 1.45) in the sample that is not restricted to those with baseline scores. These effects are, respectively, 28 and 62 percent smaller than the estimated impact of spending three years in a private school emphasized by Howell and Peterson (2002; Table 6-3).

The third set of columns in Table 6 use the race/ethnicity of the father as well as the mother to delineate the samples. That is, as in Table 5, the sample of Black (non-Hispanic) students includes children for whom *either* parent is reported as Black (non-Hispanic). The sample of Latino students likewise includes students for whom either parent is reported as Puerto Rican, Dominican, or other Hispanic. (These samples are not mutually exclusive, as is normally the case with race and ethnicity.) For students with a Black (non-Hispanic) parent, the estimated effect of attending a private school for three years is 2.3 points (t=1.04), 75 percent smaller than Peterson and Howell's estimate for students with a Black mother.

6. Conclusion

Our reanalysis of the New York City school voucher experiment suggests that the positive effect of vouchers on the achievement of African American students emphasized by previous researchers is much less robust than commonly acknowledged. Most importantly, if the wave of students who were enrolled in Kindergarten when the experiment began is included in the sample, the effect of vouchers is greatly attenuated. As the results in Table 5 indicate, treating mother and father's race symmetrically further attenuates the effect of school vouchers for African American children. The evidence is stronger that the availability of private school vouchers raised achievement on math than on reading exams after three years, but both effects are relatively small if the sample includes students with missing baseline test scores and students who have at least one Black (non-Hispanic) parent.

Below is a list of several issues about experimental program evaluation that we believe our analysis raises:

- Researchers are often unsure as to whether they should or should not control for baseline characteristics when a treatment is randomly assigned. We would advise that key results be presented *both* ways, with and without baseline characteristics (and with and without varying samples).[24] In expectation, the treatment effect should not change; if it does, more work is needed to understand why.
- Controlling for baseline characteristics can be justified if their inclusion increases the precision of the key estimates. As a practical matter, however, controlling for baseline

[24] This is often done in the academic economics literature. For example, Angrist, Bettinger, and Bloom, et al. (2003) present results with and without baseline controls in their analysis of a natural experiment involving vouchers in Colombia. Researchers are on stronger grounds for presenting results that do not condition on baseline covariates when they have an actual randomized experiment.

BEST COPY AVAILABLE

characteristics tends to reduce the sample size, which could well offset the decline in residual variance and create a non-representative sample.

- Simplicity and transparency are valuable in their own right and can help prevent mistakes. These benefits may be well worth the loss of some precision. A complicated design increases the likelihood of error down the road – for example, in the derivation of weights or in the delineation of strata within which the treatment is randomly assigned. An under appreciated virtue of presenting results *without* baseline covariates is that the results are transparent and simple, and therefore less prone to human error.
- Having a broader sample expands the population to which the results can be generalized. One wonders why the foundations that funded the voucher experiment paid the additional costs to grant scholarships and administer follow-up tests to Kindergarteners and other students with missing baseline tests if they were not to be used in the analysis.[25] Because these students make up more than 40 percent of the sample used in previous analyses, there was also a loss in efficiency from excluding them from the sample. (It is unclear whether follow-up tests were administered to students with missing baseline test scores in the Dayton and Washington experiments.)
- Researchers often make data available to other scholars, but limit the data to the sample and variables used in their previous analysis. Mathematica deserves much credit for quickly making data available (for a small fee) to outside researchers who agreed to adhere to certain confidentiality requirements, and for providing all their data, not just the subset used to generate results presented in previously published reports.

---

[25] Certainly the participating students would wonder why they were inconvenienced on three weekends to take exams that were not analyzed by the research team.

- The definition of race can be more than an incidental detail. Researchers should think carefully about the assignment of a child's race, especially if the child or parent is not asked to directly report the child's race. Because race is inherently subjective, there may be benefits from exploring the sensitivity of the findings to alternative definitions of race. Government statistics typically define race and ethnicity separately. Using a consistent and clear definition of race is necessary if results are to be compared across studies. For example, NAEP data from NCES define race to include students of any ethnicity, so the Black-White gap from NAEP is not a comparable benchmark for statistics pertaining to students with a Black (non-Hispanic) mother from the voucher experiments.
- Although we do not think that a fully specified and empirically verified theoretical model is necessary to interpret experimental results (let alone achievable), a plausible and testable theoretical explanation can help avoid mistakes in interpretation and policy. For example, the lack of a convincing theory for why African American students would benefit from vouchers while Hispanic students from the same schools would not is a cause for concern in interpreting the New York City experiment. In its simplest form, Howell and Peterson's model of constrained residential choice would predict that poor Hispanic students attending the same initial public schools as African American students would experience a rise in test scores from vouchers, at variance with the data. Combined with our finding that the effect of vouchers for African American students is more fragile than previous analyses of the data have suggested, we would counsel caution in concluding that vouchers raised achievement for African American students in New York City. The safest conclusion is probably that the provision of vouchers did not lower the scores of African American students.

References

Angrist, Joshua D., Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer, "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," forthcoming, *American Economic Review*, 2003.

Angrist, Joshua D., Guido Imbens, and Donald B. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, vol. 91, 1996, pp. 444-455.

Cochran, William G. and Gertrude M. Cox, *Experimental Designs*, New York: John Wiley and Sons, 2nd edition, 1957.

Darity, William, Darrick Hamilton and Jason Dietrich, "Passing on Blackness: Latinos, Race and Earnings in the USA," Mimeo., University of North Carolina, September 2001.

Hill, Jennifer L., Donald B. Rubin, and Neal Thomas. "The Design of the New York School Choice Scholarship Program Evaluation," in *Donald Campbell's Legacy*, edited by Leonard Bickman, Sage Publications, 2000.

Howell, William G. and Paul E. Peterson, *The Education Gap: Vouchers and Urban Schools*, Washington, DC: Brookings Institute Press, 2002 (Advance Reading Copy).

Keith, Verna M., and Cedric Herring, "Skin Tone and Stratification in the Black Community," *American Journal of Sociology*, Vol. 97, No. 3. (Nov., 1991), pp. 760-778.

Krueger, Alan, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, vol. 114, no. 2, May, 1999, pp. 497-532.

Krueger, Alan and Diane Whitmore, "Would Smaller Classes Help Close the Black-White Achievement Gap?" In *Bridging the Achievement Gap*, edited by John Chubb and Tom Loveless, Washington, DC: Brookings Institute Press, 2002.

Mathematica Policy Research, Inc., Press Release, "Voucher Claims of Success Are Premature in New York City," Washington, D.C., September 2000.

Mayer, Daniel P., David E. Myers, and Christina Clark Tuttle, "Appendix E: African American and Latino Test-Score Differences," in "School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program," edited by Daniel P. Mayer, et al.,Washington, DC: Mathematica Policy Research, Inc., 2002.

Mayer, Daniel P., Paul E. Peterson, David E. Myers, Christina Clark Tuttle, and William G. Howell, "School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program," Washington, DC: Mathematica Policy Research, Inc., 2002.

Moulton, Brent R., "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics*, vol. 72, no. 2, May 1990, pp. 334-338.

Neal, Derek, "How Vouchers Could Change the Market for Education," forthcoming *Journal of Economic Perspectives*, 2002.

Peterson, Paul E., David Myers and William G. Howell, "An Evaluation of the New York City Scholarships Program: The First Year," Washington, DC: Mathematica Policy Research, Inc., 1998.

Rouse, Cecilia Elena, "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," NBER Working Paper Number 5964 (March, 1997).

Scarr, Sandra, Andrew J. Pakstis, Solomon Katz and William Barker, "Absence of a relationship between degree of white ancestry and intellectual skills within a black population," *Human Genetics* 39, 1977, pp. 69-86.

Table 1: Treatment-Control Differences at Baseline by Assignment Status, Selected Variables

Group	Mean treatment	Mean control	Raw T-C Difference	t-ratio	Conditional on Lottery Strata: T-C Difference	t-ratio	Sample size
MOTHER'S YEARS OF EDUCATION							
Overall	12.84	12.78	0.07	0.61	0.07	0.69	2,476
Black- all	13.03	12.96	0.07	0.48	0.11	0.77	1,081
Black-ch0	12.83	12.92	-0.09	-0.32	-0.10	-0.33	233
Black-ch1	13.10	13.22	-0.13	-0.47	-0.04	-0.15	215
Black-ch2	13.22	13.08	0.14	0.52	0.18	0.62	247
Black-ch3	12.84	12.88	-0.04	-0.17	0.00	-0.01	216
Black-ch4	13.19	12.57	0.62	1.95	0.55	1.52	169
Latino-all	12.46	12.41	0.06	0.37	0.08	0.50	1,186
MOTHER GRADUATED COLLEGE							
Overall	0.11	0.09	0.02	0.98	0.01	0.90	2,476
Black- all	0.12	0.09	0.03	1.31	0.04	1.73	1,081
Black-ch0	0.12	0.05	0.07	1.80	0.06	1.50	233
Black-ch1	0.07	0.12	-0.05	-1.12	-0.01	-0.30	215
Black-ch2	0.17	0.13	0.03	0.63	0.04	0.68	247
Black-ch3	0.07	0.08	-0.01	-0.28	0.00	-0.10	216
Black-ch4	**0.20**	**0.03**	**0.16**	**2.95**	**0.13**	**2.29**	**169**
Latino-all	0.06	0.07	0.00	-0.18	0.00	-0.13	1,186
ANNUAL INCOME							
Overall	$10,291	$9,982	$308	0.81	$298	0.78	2,433
Black- all	10,629	10,283	345	0.57	222	0.36	1,068
Black-ch0	11,933	10,306	1,627	1.30	1,266	0.92	228
Black-ch1	8,736	10,433	-1,696	-1.51	-1,357	-1.07	213
Black-ch2	10,669	9,989	680	0.68	471	0.43	246
Black-ch3	9,846	10,750	-904	-0.87	-981	-0.89	211
Black-ch4	**12,438**	**9,890**	**2,548**	**2.19**	2,602	1.84	169
Latino-all	9,941	9,470	471	0.88	347	0.67	1,161
BASELINE TEST SCORES							
COMPOSITE MATH AND READING SCORE (AVERAGE NPR)							
Overall	19.97	20.83	-0.86	-0.84	-0.66	-0.63	1,851
Black- all	20.11	20.41	-0.30	-0.21	-0.24	-0.17	806
Black-ch0	NA	NA	NA	NA	NA	NA	NA
Black-ch1	17.50	21.84	-4.35	-1.53	-4.22	-1.32	203
Black-ch2	23.35	22.45	0.90	0.34	-0.27	-0.09	230
Black-ch3	16.15	18.04	-1.89	-0.77	-2.01	-0.79	214
Black-ch4	**24.82**	**18.23**	**6.59**	**2.17**	5.79	1.81	159
Latino-all	18.02	20.01	-1.99	-1.40	-1.81	-1.18	876
READING (NPR)							
Overall	22.88	24.55	-1.67	-1.37	-1.42	-1.15	1,851
Black- all	24.40	25.35	-0.95	-0.54	-0.76	-0.42	806
Black-ch0	NA	NA	NA	NA	NA	NA	NA
Black-ch1	24.71	31.27	-6.56	-1.64	-7.10	-1.52	203
Black-ch2	26.44	25.96	0.48	0.14	0.82	0.21	230
Black-ch3	19.72	21.69	-1.97	-0.72	-2.92	-1.04	214
Black-ch4	**27.73**	**20.58**	**7.14**	**2.16**	**7.65**	**2.11**	**159**
Latino-all	19.76	23.33	-3.57	-2.07	-3.17	-1.70	876

MATH (NPR)							
Overall	17.06	17.11	-0.04	-0.04	0.11	0.09	1,851
Black- all	15.82	15.46	0.36	0.25	0.28	0.19	806
Black-ch0	NA	NA	NA	NA	NA	NA	NA
Black-ch1	10.28	12.42	-2.14	-0.95	-1.33	-0.56	203
Black-ch2	20.26	18.93	1.32	0.51	-1.35	-0.48	230
Black-ch3	12.58	14.40	-1.82	-0.70	-1.10	-0.40	214
Black-ch4	21.92	15.89	6.04	1.66	3.93	1.12	159
Latino-all	16.28	16.70	-0.42	-0.28	-0.45	-0.29	876
MOTHER EMPLOYED FULL TIME							
Overall	0.23	0.22	0.00	0.21	0.01	0.25	2,479
Black- all	0.28	0.26	0.03	0.72	0.02	0.45	1,083
Black-ch0	0.36	0.27	0.09	1.39	0.04	0.63	233
Black-ch1	0.19	0.29	-0.10	-1.57	-0.08	-1.09	216
Black-ch2	0.27	0.20	0.06	1.02	0.07	1.20	248
Black-ch3	0.25	0.31	-0.06	-0.92	-0.09	-1.27	216
Black-ch4	**0.36**	**0.19**	**0.17**	**2.38**	**0.17**	**2.08**	**169**
Latino-all	0.20	0.19	0.00	0.10	0.01	0.17	1,187
STUDENT SEX (MALE)							
Overall	0.50	0.48	0.01	0.69	0.01	0.65	2,617
Black- all	0.47	0.53	-0.06	-1.75	-0.06	-1.80	1,134
Black-ch0	**0.41**	**0.60**	**-0.19**	**-2.69**	**-0.22**	**-3.29**	**241**
Black-ch1	0.49	0.58	-0.09	-1.22	-0.15	-1.74	225
Black-ch2	0.51	0.54	-0.03	-0.46	-0.09	-1.27	256
Black-ch3	0.46	0.43	0.03	0.46	0.00	0.01	232
Black-ch4	0.48	0.45	0.03	0.33	0.05	0.57	180
Latino-all	0.49	0.44	0.05	1.59	0.05	1.46	1,253
MOTHER'S PLACE OF BIRTH USA							
Overall	0.52	0.55	-0.03	-1.03	-0.03	-1.15	2,593
Black- all	**0.75**	**0.82**	**-0.07**	**-1.98**	**-0.08**	**-2.55**	**1,132**
Black-ch0	0.74	0.77	-0.03	-0.47	0.00	-0.07	240
Black-ch1	0.76	0.86	-0.10	-1.65	**-0.15**	**-2.37**	**225**
Black-ch2	0.77	0.78	-0.02	-0.29	-0.05	-0.92	261
Black-ch3	0.82	0.82	-0.01	-0.12	-0.05	-0.94	227
Black-ch4	**0.64**	**0.87**	**-0.23**	**-3.42**	**-0.22**	**-3.02**	**178**
Latino-all	0.31	0.36	-0.04	-1.21	-0.05	-1.41	1,245
FOOD STAMP RECIPIENT							
Overall	0.65	0.69	-0.03	-1.26	-0.03	-1.11	2,474
Black- all	0.63	0.70	-0.06	-1.74	-0.05	-1.20	1,074
Black-ch0	0.61	0.70	-0.09	-1.34	-0.08	-1.09	227
Black-ch1	0.71	0.68	0.02	0.33	0.06	0.78	216
Black-ch2	0.59	0.70	-0.11	-1.65	-0.14	-1.87	248
Black-ch3	0.69	0.68	0.01	0.18	0.00	0.05	217
Black-ch4	**0.54**	**0.73**	**-0.19**	**-2.35**	**-0.21**	**-2.32**	**165**
Latino-all	0.67	0.70	-0.03	-0.75	-0.02	-0.62	1,190
AFDC RECIPIENT							
Overall	0.57	0.60	-0.04	-1.33	-0.03	-1.11	2,375
Black- all	0.58	0.65	-0.07	-1.81	-0.06	-1.40	1,030
Black-ch0	0.57	0.64	-0.07	-0.95	-0.07	-0.92	218
Black-ch1	0.64	0.63	0.01	0.09	0.03	0.31	204
Black-ch2	**0.51**	**0.67**	**-0.16**	**-2.36**	**-0.18**	**-2.44**	**241**
Black-ch3	0.64	0.60	0.04	0.56	0.03	0.32	206
Black-ch4	**0.53**	**0.73**	**-0.20**	**-2.59**	**-0.24**	**-2.82**	**160**
Latino-all	0.54	0.59	-0.05	-1.34	-0.05	-1.27	1,143

MEDICAID RECIPIENT							
Overall	**0.62**	**0.69**	**-0.07**	**-2.53**	**-0.06**	**-2.29**	**2,299**
Black- all	**0.59**	**0.69**	**-0.09**	**-2.31**	-0.07	-1.86	**1,003**
Black-ch0	0.61	0.69	-0.08	-1.14	-0.07	-0.91	217
Black-ch1	0.57	0.60	-0.04	-0.42	-0.01	-0.16	195
Black-ch2	**0.58**	**0.72**	**-0.14**	**-2.03**	**-0.15**	**-1.99**	**239**
Black-ch3	0.67	0.69	-0.02	-0.21	-0.02	-0.26	203
Black-ch4	**0.52**	**0.74**	**-0.22**	**-2.70**	**-0.23**	**-2.43**	**148**
Latino-all	0.62	0.70	-0.08	-2.01	-0.07	-1.79	1,098
EVER ATTENDED GIFTED STUDENT CLASSES							
Overall	0.12	0.12	0.00	-0.09	0.00	-0.08	2,547
Black- all	0.15	0.13	0.01	0.60	0.01	0.49	1,115
Black-ch0	0.08	0.09	-0.02	-0.34	0.00	-0.10	239
Black-ch1	0.12	0.11	0.01	0.17	-0.01	-0.17	222
Black-ch2	0.21	0.17	0.04	0.61	0.00	-0.01	254
Black-ch3	0.14	0.11	0.03	0.55	0.06	1.37	227
Black-ch4	0.21	0.19	0.01	0.16	0.01	0.13	173
Latino-all	0.09	0.09	-0.01	-0.29	-0.01	-0.29	1,208
EVER RECEIVED SPECIAL EDUCATION SERVICES							
Overall	0.11	0.10	0.01	0.71	0.01	0.65	2,574
Black- all	0.11	0.10	0.01	0.67	0.02	0.75	1,111
Black-ch0	0.08	0.09	-0.02	-0.37	-0.02	-0.39	238
Black-ch1	0.07	0.10	-0.03	-0.71	-0.03	-0.58	222
Black-ch2	0.13	0.10	0.03	0.73	-0.01	-0.13	252
Black-ch3	**0.16**	**0.06**	**0.09**	**2.11**	0.08	1.88	**226**
Black-ch4	0.14	0.15	-0.01	-0.20	0.03	0.46	173
Latino-all	0.11	0.12	-0.01	-0.58	-0.01	-0.63	1,239
ADDENDUM: THIRD YEAR FOLLOW-UP TEST RESULTS							
COMPOSITE MATH AND READING SCORE (AVERAGE NPR)							
Overall	26.72	27.12	-0.40	-0.34	-0.41	-0.34	1,801
Black- all	25.50	23.67	1.84	1.11	2.78	1.65	733
Black-ch0	21.62	27.13	-5.51	-1.47	-6.24	-1.58	156
Black-ch1	26.40	23.33	3.07	0.77	1.82	0.40	139
Black-ch2	23.10	21.11	1.99	0.68	2.76	0.82	177
Black-ch3	27.99	26.66	1.33	0.39	1.57	0.40	139
Black-ch4	**29.04**	**20.14**	**8.90**	**2.25**	**10.49**	**2.60**	**122**
Latino-all	26.42	28.18	-1.77	-1.06	-1.46	-0.86	932

Notes: ch-0 refers to Kindergartners at baseline, ch-1 to first graders at baseline, and so on. Observations with missing values for a particular variable are dropped from the sample in the relevant row. Bootstrap standard errors used to compute the t-ratios account for dependent observations within families. Lottery strata are the 30 groups students were placed in at the time of random assignment.

A bold font indicates that the absolute t-ratio for the within-strata treatment-control difference exceeds 1.96.

Table 2: Efficiency Comparison of Two Methods of Random Assignment
Estimated Treatment Effect and Standard Error By Method of Random Assignment

A. All Students

Year	Model	Propensity Score Match Subsample			Stratified Blocks Subsample			Relative Std. Error	Sample-Size-Adjusted Relative Std. Error
		Coeff.	Std. Error	Nobs.	Coeff.	Std. Error	Nobs.		
Year 1	Control for baseline	0.41	1.40	721	2.08	1.38	734	0.986	0.995
	Omit baseline	0.42	1.40	1048	-1.12	1.58	1032	1.129	1.120
Year 2	Control for baseline	0.55	1.49	603	0.21	1.73	596	1.161	1.154
	Omit baseline	0.24	1.59	908	-0.14	1.84	846	1.157	1.117
Year 3	Control for baseline	0.88	1.56	613	0.69	1.74	637	1.115	1.137
	Omit baseline	-0.32	1.58	900	-0.50	1.80	901	1.139	1.140
B. African American Students									
Year 1	Control for baseline	0.23	1.87	302	8.08	1.72	321	0.920	0.948
	Omit baseline	2.44	1.97	436	3.21	2.07	447	1.051	1.064
Year 2	Control for baseline	2.82	2.16	244	3.53	2.60	253	1.204	1.226
	Omit baseline	4.10	2.28	360	1.61	2.54	362	1.114	1.117
Year 3	Control for baseline	4.18	2.51	247	6.55	2.09	272	0.833	0.874
	Omit baseline	2.46	2.48	347	3.10	2.32	386	0.935	0.987

Notes: Treatment effect coefficient is from a regression of test scores on a dummy indicating assignment to receive a voucher (1=yes), original 30 lottery randomization strata dummies, and in some models baseline test scores. Bootstrap standard errors account for within-family correlation in residuals. Year 1, 2, or 3 refers to follow-up year.

Table 3a: Estimated treatment effects, with and without controlling for baseline scores

Controls for *original* 30 strata used to assign students to random assignment blocks

Test	Group	SubSample with Baseline Scores; Controls For Baseline Scores				SubSample with Baseline Scores; Omits Baseline Scores				Full Sample; Omits Baseline Scores			
First Follow-up Test:		NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio
Composite	Overall	1,455	1.30	0.96	1.35	1,455	0.40	1.29	0.31	2,080	-0.36	1.06	-0.34
	Black	623	**4.20**	**1.28**	**3.28**	623	**3.89**	**1.73**	**2.26**	883	**2.83**	**1.41**	**2.01**
	Latino	709	-0.38	1.37	-0.28	709	-1.96	1.67	-1.17	1,021	-1.63	1.48	-1.11
Reading	Overall	1,455	1.15	1.04	1.11	1,455	0.04	1.37	0.03	2,080	-1.13	1.17	-0.96
	Black	623	**3.35**	**1.57**	**2.14**	623	2.90	1.99	1.46	883	1.64	1.72	0.95
	Latino	709	-0.55	1.44	-0.38	709	-2.46	1.80	-1.36	1,021	-2.34	1.66	-1.41
Math	Overall	1,455	1.45	1.18	1.23	1,455	0.76	1.45	0.52	2,080	0.42	1.17	0.36
	Black	623	**5.04**	**1.53**	**3.30**	623	**4.88**	**1.91**	**2.55**	883	**4.02**	**1.51**	**2.67**
	Latino	709	-0.22	1.69	-0.13	709	-1.46	1.90	-0.77	1,021	-0.93	1.60	-0.58
Second Follow-up Test:													
Composite	Overall	1,199	0.40	1.14	0.36	1,199	0.11	1.44	0.08	1,754	0.05	1.22	0.04
	Black	497	3.14	1.68	1.87	497	3.60	1.95	1.84	722	2.86	1.70	1.68
	Latino	612	-0.60	1.52	-0.40	612	-2.18	1.97	-1.11	902	-1.84	1.63	-1.13
Reading	Overall	1,199	1.31	1.11	1.17	1,199	0.82	1.46	0.56	1,754	0.63	1.30	0.48
	Black	497	**3.49**	**1.77**	**1.98**	497	3.81	2.11	1.81	722	2.87	1.92	1.49
	Latino	612	0.16	1.56	0.10	612	-1.61	2.01	-0.80	902	-1.25	1.72	-0.73
Math	Overall	1,199	-0.50	1.52	-0.33	1,199	-0.59	1.73	-0.34	1,754	-0.53	1.39	-0.38
	Black	497	2.79	2.18	1.28	497	3.39	2.36	1.44	722	2.85	1.96	1.46
	Latino	612	-1.36	1.94	-0.70	612	-2.76	2.31	-1.20	902	-2.44	1.84	-1.32
Third Follow-up Test													
Composite	Overall	1,250	0.84	1.14	0.74	1,250	0.33	1.40	0.24	1,801	-0.41	1.20	-0.34
	Black	519	**5.21**	**1.62**	**3.21**	519	**5.14**	**1.93**	**2.66**	733	2.78	1.68	1.65
	Latino	637	-1.11	1.61	-0.69	637	-1.67	1.99	-0.84	932	-1.46	1.70	-0.86
Reading	Overall	1,250	0.29	1.26	0.23	1,250	-0.36	1.52	-0.24	1,801	-0.96	1.26	-0.76
	Black	519	**3.73**	**1.89**	**1.98**	519	3.47	2.21	1.57	733	1.51	1.84	0.82
	Latino	637	-1.91	1.78	-1.07	637	-2.63	2.13	-1.24	932	-1.97	1.83	-1.08
Math	Overall	1,250	1.40	1.30	1.08	1,250	1.02	1.53	0.67	1,801	0.13	1.34	0.10
	Black	519	**6.68**	**1.84**	**3.64**	519	**6.81**	**2.08**	**3.27**	733	**4.06**	**1.86**	**2.18**
	Latino	637	-0.30	1.84	-0.16	637	-0.72	2.20	-0.38	932	-0.96	1.88	-0.51

Notes: Dependent variable is test score NPR. Reported coefficient is coefficient on voucher offer dummy. All regressions control for original 30 randomization strata. Bootstrap standard errors are robust to correlation in residuals among students in the same family. Bold font indicates that the absolute t-ratio exceeds 1.96.

Table 3b: Estimated treatment effects, with and without controlling for baseline scores

Controls for 30 randomization strata as defined by Mayer, Peterson and Myers, et al. (2000)

Test	Group	SubSample with Baseline Scores; Controls For Baseline Scores				SubSample with Baseline Scores; Omits Baseline Scores				Full Sample; Omits Baseline Scores			
First Follow-up Test:		NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio
Composite	Overall	1,455	1.20	0.97	1.24	1,455	0.23	1.30	0.18	2,080	-0.40	1.06	-0.38
	Black	623	**4.43**	**1.27**	**3.48**	623	**3.87**	**1.73**	**2.24**	883	2.64	1.42	1.86
	Latino	709	-0.66	1.38	-0.48	709	-2.59	1.77	-1.46	1,021	-2.05	1.56	-1.31
Reading	Overall	1,455	1.01	1.05	0.96	1,455	-0.18	1.38	-0.13	2,080	-1.17	1.18	-0.99
	Black	623	**3.47**	**1.57**	**2.21**	623	2.74	1.98	1.38	883	1.46	1.72	0.85
	Latino	709	-0.70	1.45	-0.48	709	-3.02	1.88	-1.60	1,021	-2.60	1.72	-1.51
Math	Overall	1,455	1.39	1.17	1.19	1,455	0.65	1.45	0.45	2,080	0.37	1.16	0.31
	Black	623	**5.39**	**1.52**	**3.56**	623	**5.01**	**1.92**	**2.60**	883	**3.82**	**1.54**	**2.48**
	Latino	709	-0.63	1.70	-0.37	709	-2.17	2.00	-1.08	1,021	-1.50	1.67	-0.90
Second Follow-up Test:													
Composite	Overall	1,199	0.46	1.13	0.41	1,199	0.14	1.44	0.10	1,754	0.05	1.22	0.04
	Black	497	3.27	1.68	1.95	497	3.51	1.94	1.81	722	2.58	1.70	1.52
	Latino	612	-0.60	1.50	-0.40	612	-2.34	1.97	-1.19	902	-1.83	1.62	-1.13
Reading	Overall	1,199	1.35	1.11	1.22	1,199	0.83	1.45	0.57	1,754	0.62	1.30	0.48
	Black	497	**3.44**	**1.74**	**1.97**	497	3.50	2.06	1.70	722	2.52	1.90	1.32
	Latino	612	0.17	1.55	0.11	612	-1.75	2.05	-0.85	902	-1.17	1.73	-0.68
Math	Overall	1,199	-0.43	1.51	-0.28	1,199	-0.55	1.73	-0.32	1,754	-0.52	1.39	-0.38
	Black	497	3.10	2.18	1.42	497	3.53	2.36	1.49	722	2.64	1.97	1.35
	Latino	612	-1.37	1.92	-0.72	612	-2.93	2.29	-1.28	902	-2.48	1.82	-1.36
Third Follow-up Test													
Composite	Overall	1,250	0.93	1.13	0.82	1,250	0.24	1.40	0.17	1,801	-0.31	1.18	-0.26
	Black	519	**5.50**	**1.61**	**3.42**	519	**5.30**	**1.91**	**2.77**	733	2.87	1.69	1.71
	Latino	637	-0.95	1.57	-0.60	637	-1.86	2.00	-0.93	932	-1.35	1.66	-0.81
Reading	Overall	1,250	0.27	1.25	0.21	1,250	-0.56	1.52	-0.37	1,801	-0.95	1.25	-0.77
	Black	519	**3.97**	**1.87**	**2.13**	519	3.56	2.18	1.63	733	1.53	1.83	0.84
	Latino	637	-1.85	1.73	-1.07	637	-2.90	2.15	-1.35	932	-1.81	1.78	-1.01
Math	Overall	1,250	1.59	1.28	1.24	1,250	1.05	1.52	0.69	1,801	0.33	1.32	0.25
	Black	519	**7.03**	**1.82**	**3.86**	519	**7.04**	**2.06**	**3.41**	733	**4.22**	**1.85**	**2.28**
	Latino	637	-0.05	1.82	-0.03	637	-0.82	2.19	-0.38	932	-0.89	1.83	-0.48

Notes: Dependent variable is test score NPR. Reported coefficient is coefficient on voucher offer dummy. All regressions control for 30 revised randomization strata. Bootstrap standard errors are robust to correlation in residuals among students in the same family. Bold font indicates that the absolute t-ratio exceeds 1.96.

Table 4: Estimated treatment effects, with varying controls for baseline characteristics

Test	Group	SubSample with Baseline Scores Controls for baseline scores, cohort & other covariates				Full Sample Controls for cohort & other covariates (omits baseline test scores)				Full Sample Controls for cohort, baseline scores & interaction for missing scores			Full Sample Controls for cohort, baseline scores, interaction for missing scores & other covariates		
First Follow-up Test:		NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio	Coefficient	S.E.	t-ratio	Coefficient	S.E.	t-ratio
Composite	Overall	1,455	1.73	0.92	1.88	2,080	-0.10	0.98	-0.10	0.51	0.86	0.59	0.61	0.84	0.73
	Black	623	**3.96**	**1.33**	**2.98**	883	**2.87**	**1.38**	**2.08**	**2.95**	**1.16**	**2.55**	**2.93**	**1.20**	**2.45**
	Latino	709	-0.21	1.35	-0.16	1,021	-1.86	1.37	-1.36	-0.14	1.22	-0.12	-0.44	1.20	-0.37
Reading	Overall	1,455	1.33	1.02	1.31	2,080	-0.95	1.10	-0.87	-0.12	0.98	-0.13	-0.12	0.96	-0.13
	Black	623	2.55	1.62	1.57	883	1.36	1.67	0.82	1.79	1.49	1.20	1.54	1.50	1.02
	Latino	709	-0.48	1.41	-0.34	1,021	-2.30	1.59	-1.44	-0.43	1.39	-0.31	-0.69	1.40	-0.49
Math	Overall	1,455	2.14	1.14	1.88	2,080	0.75	1.10	0.68	1.13	0.98	1.16	1.34	0.97	1.38
	Black	623	**5.37**	**1.55**	**3.46**	883	**4.38**	**1.49**	**2.94**	**4.12**	**1.27**	**3.23**	**4.33**	**1.32**	**3.27**
	Latino	709	0.05	1.74	0.03	1,021	-1.43	1.48	-0.97	0.15	1.37	0.11	-0.19	1.36	-0.14
Second Follow-up Test:															
Composite	Overall	1,199	0.61	1.12	0.54	1,754	-0.38	1.15	-0.33	0.29	1.04	0.28	-0.09	1.00	-0.09
	Black	497	**3.53**	**1.67**	**2.11**	722	2.98	1.67	1.78	2.61	1.57	1.67	2.59	1.52	1.71
	Latino	612	-1.24	1.51	-0.82	902	-2.37	1.49	-1.60	-0.92	1.32	-0.70	-1.34	1.27	-1.05
Reading	Overall	1,199	1.66	1.13	1.47	1,754	0.36	1.23	0.29	1.00	1.11	0.90	0.73	1.08	0.68
	Black	497	**3.99**	**1.87**	**2.14**	722	2.99	1.88	1.60	2.85	1.74	1.64	2.71	1.72	1.57
	Latino	612	-0.14	1.60	-0.09	902	-1.41	1.64	-0.86	-0.19	1.44	-0.13	-0.30	1.43	-0.21
Math	Overall	1,199	-0.45	1.47	-0.31	1,754	-1.13	1.32	-0.85	-0.43	1.23	-0.34	-0.92	1.20	-0.76
	Black	497	3.08	2.08	1.48	722	2.97	1.94	1.53	2.37	1.87	1.27	2.48	1.81	1.37
	Latino	612	-2.34	1.95	-1.20	902	**-3.33**	**1.67**	**-2.00**	-1.64	1.55	-1.06	-2.38	1.49	-1.60
Third Follow-up Test															
Composite	Overall	1,250	1.10	1.08	1.01	1,801	-0.48	1.14	-0.42	0.11	1.04	0.11	-0.04	1.00	-0.04
	Black	519	**4.50**	**1.62**	**2.78**	733	2.12	1.68	1.27	2.80	1.53	1.83	2.19	1.51	1.45
	Latino	637	-1.01	1.57	-0.64	932	-1.31	1.62	-0.81	-0.72	1.43	-0.50	-0.68	1.42	-0.48
Reading	Overall	1,250	0.50	1.23	0.41	1,801	-0.99	1.22	-0.82	-0.30	1.10	-0.27	-0.51	1.08	-0.47
	Black	519	3.20	1.90	1.68	733	0.84	1.82	0.46	1.70	1.67	1.02	1.05	1.65	0.64
	Latino	637	-1.91	1.75	-1.10	932	-1.63	1.78	-0.92	-1.02	1.57	-0.65	-0.97	1.58	-0.61
Math	Overall	1,250	1.69	1.25	1.36	1,801	0.04	1.27	0.03	0.52	1.20	0.43	0.42	1.16	0.36
	Black	519	**5.81**	**1.86**	**3.12**	733	3.41	1.88	1.81	**3.90**	**1.74**	**2.23**	3.32	1.74	1.91
	Latino	637	-0.11	1.82	-0.06	932	-0.98	1.77	-0.55	-0.41	1.63	-0.25	-0.39	1.62	-0.24

Notes: Dependent variable is test score NPR. Reported coefficient is coefficient on voucher offer dummy. All regressions control for original 30 randomization strata and cohort (initial grade) dummies. The first set of columns also control for baseline test scores. The last two sets of columns control for baseline test scores and a dummy indicating whether baseline scores are missing and an interaction between that dummy and baseline test scores. When indicated, "other covariates" included are gender, mother's education, mother's full-time or part-time employment status, special education, gifted or talented class, welfare, log family income, student's age, mother's place of birth, English spoken at home, mother in current residence less than one year, mother's religion Catholic, and dummies indicating whether the covariates are missing.

Bootstrap standard errors account for dependent observations among students in the same family. Bold font indicates absolute t-ratio in excess of 1.96.

BEST COPY AVAILABLE

Table 5: Estimated treatment effects for all Black students

Sample includes students whose mother or father is Black/African American (non-Hispanic)

	SubSample w/ Baseline Test Controls for baseline scores & covariates				Full Sample Controls for covariates except baseline scores				Full Sample Controls for covariates, baseline scores & interaction for missing scores			
Test/Year	N	Coefficient	S.E.	t-ratio	N	Coefficient	S.E.	t-ratio	N	Coefficient	S.E.	t-ratio
Year One												
Composite	684	**3.52**	**1.27**	**2.76**	968	1.93	1.38	1.40	968	2.19	1.14	1.92
Reading	684	2.06	1.53	1.35	968	0.29	1.65	0.18	968	0.61	1.43	0.43
Math	684	**4.98**	**1.53**	**3.25**	968	**3.57**	**1.48**	**2.41**	968	**3.77**	**1.28**	**2.94**
Year Two												
Composite	545	2.34	1.62	1.44	787	1.83	1.63	1.12	787	1.69	1.44	1.17
Reading	545	3.05	1.77	1.32	787	1.97	1.83	1.08	787	1.86	1.65	1.13
Math	545	1.63	2.03	0.80	787	1.68	1.86	0.90	787	1.52	1.70	0.89
Year Three												
Composite	572	**4.10**	**1.61**	**2.54**	807	1.24	1.66	0.75	807	1.52	1.46	1.04
Reading	572	3.24	1.88	1.72	807	0.20	1.79	0.11	807	0.57	1.61	0.36
Math	572	**4.97**	**1.87**	**2.66**	807	2.27	1.86	1.22	807	2.46	1.69	1.46

Notes: Dependent variable is test score NPR. Reported coefficient is coefficient on voucher offer dummy. All regressions control for 30 randomization strata, cohort dummies, gender, mother's education, mother's full-time or part-time employment status, special education, gifted or talented class, welfare, log family income, student's age, mother's place of birth, English spoken at home, mother in current residence less than one year, and mother's religion. The first set of columns also control for baseline test scores. The last set of columns also control for a dummy indicating baseline scores missing and an interaction between that dummy and baseline test scores.

Bootstrap standard errors account for dependent observations among students in the same family. Bold font indicates absolute t-ratio in excess of 1.96.

Table 6: Two-Stage Least Squares estimates of effect of years in private school, with varying controls for baseline covariates

Test	Group	SubSample with Baseline Scores Controls for baseline scores & covariates			Full Sample; Mother's Race/Ethnicity Controls for covariates, baseline test scores & interaction for missing scores			Full Sample; Either Parent's Race/Ethnicity Controls for covariates, baseline test scores & interaction for missing scores		
		NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio
First Follow-up Test:													
Composite	Overall	1,449	2.36	1.26	1.88	2,080	0.83	1.20	0.69	2,080	0.83	1.20	0.69
	Black	622	**5.14**	**1.75**	**2.93**	883	**3.92**	**1.63**	**2.40**	968	2.92	1.53	1.90
	Latino	704	-0.34	1.92	-0.18	1,021	-0.65	1.76	-0.37	1,117	-0.22	1.75	-0.13
Reading	Overall	1,449	1.82	1.37	1.34	2,080	-0.19	1.38	-0.14	2,080	-0.19	1.38	-0.14
	Black	622	3.31	2.10	1.58	883	2.07	2.02	1.02	968	0.82	1.91	0.43
	Latino	704	-0.67	2.00	-0.33	1,021	-1.07	2.03	-0.53	1,117	-1.05	2.02	-0.52
Math	Overall	1,449	2.89	1.58	1.83	2,080	1.85	1.39	1.33	2,080	1.85	1.39	1.33
	Black	622	**6.96**	**2.06**	**3.37**	883	**5.78**	**1.81**	**3.20**	968	**5.02**	**1.74**	**2.88**
	Latino	704	-0.02	2.46	-0.01	1,021	-0.24	1.98	-0.12	1,117	0.60	1.96	0.31
Second Follow-up Test:													
Composite	Overall	1,199	0.42	0.77	0.54	1,754	-0.07	0.72	-0.09	1,754	-0.07	0.72	-0.09
	Black	497	**2.31**	**1.09**	**2.12**	722	1.81	1.06	1.71	787	1.18	1.00	1.18
	Latino	612	-0.89	1.07	-0.83	902	-0.96	0.91	-1.05	984	-0.88	0.93	-0.94
Reading	Overall	1,199	1.15	0.78	1.47	1,754	0.52	0.78	0.67	1,754	0.52	0.78	0.67
	Black	497	**2.60**	**1.22**	**2.14**	722	1.89	1.20	1.57	787	1.30	1.14	1.13
	Latino	612	-0.10	1.14	-0.09	902	-0.21	1.03	-0.21	984	-0.72	1.05	-0.68
Math	Overall	1,199	-0.31	1.01	-0.31	1,754	-0.66	0.86	-0.76	1,754	-0.66	0.86	-0.76
	Black	497	2.01	1.35	1.49	722	1.73	1.26	1.38	787	1.06	1.18	0.90
	Latino	612	-1.68	1.38	-1.22	902	-1.71	1.07	-1.60	984	-1.05	1.10	-0.95
Third Follow-up Test													
Composite	Overall	1,250	0.55	0.54	1.01	1,801	-0.02	0.53	-0.04	1,801	-0.02	0.53	-0.04
	Black	519	**2.23**	**0.81**	**2.75**	733	1.15	0.79	1.45	807	0.78	0.75	1.04
	Latino	637	-0.52	0.79	-0.65	932	-0.35	0.73	-0.48	1,015	-0.20	0.71	-0.28
Reading	Overall	1,250	0.25	0.61	0.41	1,801	-0.27	0.57	-0.47	1,801	-0.27	0.57	-0.47
	Black	519	1.58	0.94	1.69	733	0.55	0.86	0.64	807	0.30	0.83	0.36
	Latino	637	-0.98	0.89	-1.10	932	-0.50	0.82	-0.61	1,015	-0.28	0.79	-0.36
Math	Overall	1,250	0.84	0.62	1.35	1,801	0.22	0.61	0.36	1,801	0.22	0.61	0.36
	Black	519	**2.88**	**0.94**	**3.05**	733	1.76	0.92	1.91	807	1.27	0.87	1.46
	Latino	637	-0.05	0.92	-0.06	932	-0.20	0.84	-0.24	1,015	-0.12	0.80	-0.15

Notes: Dependent variable is test score NPR. Reported coefficient is coefficient on years in private school. All regressions control for 30 randomization strata, cohort dummies, gender, mother's education, mother's full-time or part-time employment status, special education, gifted or talented class, welfare, log family income, student's age, mother's place of birth, English spoken at home, mother in current residence less than one year, mother's religion, and dummies indicating whether the covariates are missing. The first set of columns also control for baseline test scores. The last two sets of columns also control for a dummy indicating whether baseline scores are missing and an interaction between that dummy and baseline test scores.

Bootstrap standard errors account for dependent observations among students in the same family. Bold font indicates absolute t-ratio in excess of 1.96.

BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Another Look at the New York City School Voucher Experiment	EA 032 339
Author(s): Alan B. Krueger, Pei Zhu	
Corporate Source:	Publication Date: Jan 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Alan Krueger	Printed Name/Position/Title: Alan Krueger
Organization/Address: Princeton Univ.	Telephone: (609)-258-4048 FAX:
	E-Mail Address: Date: 2/25/03

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Educational Management
5207 University of Oregon
Eugene OR 97403-5207
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com